What Is Claimed Is:

1   1.   A method for learning a generative model for text, comprising:

2   receiving a current model, which contains terminal nodes representing

3   random variables for words and can contain cluster nodes representing clusters of

4   conceptually related words;

5   wherein nodes in the current model are coupled together by weighted

6   links, so that if a cluster node in the probabilistic model fires, a weighted link

7   from the cluster node to another node causes the other node to fire with a

8   probability proportionate to the link weight;

9   receiving a set of training documents, wherein each training document

10   contains a set of words; and

11   applying the set of training documents to the current model to produce a

12   new model.

1   2.   The method of claim 1, wherein applying the set of training

2   documents to the current model involves:

3   applying the set of training documents to the links defined in the current

4   model to produce functions for weights for corresponding links in the new model;

5   and

6   optimizing the functions to produce weights for links in the new model.

1   3.   The method of claim 2, wherein for a given link, producing

2   functions for a weight on the given link involves:

3   producing a function for the given link for each document in the set of

4   training documents; and

71

5   multiplying the functions for each document together to produce a

6 function to be optimized for the given link.

1   4. The method of claim 3, wherein for the given link the function for

2 a document is an approximation of the probability of the document's terminals

3 firing as a function of the weight on the given link, keeping all other link weights

4 in the model constant.

1   5. The method of claim 1, wherein the method further comprises

2 iteratively:

3   considering the new model to be the current model; and

4   applying training documents to the current model to produce a subsequent

5 new model.

1   6. The method of claim 5, wherein during an initial iteration, the

2 method further comprises generating an initial current model from a set of words

3 by:

4   generating a universal node that is always active;

5   generating terminal nodes representing words in the set of words; and

6   directly linking the universal node to the terminal nodes.

1   7. The method of claim 5, wherein each iteration uses twice as many

2 training documents as the previous iteration until all available training documents

3 are used.

1   8.   The method of claim 1, wherein producing the new model

2 additionally involves selectively introducing new links from clusters to nodes and

3 from clusters to clusters.

1   9.   The method of claim 8, wherein introducing a new link can

2 involve:

3   considering a cluster that is likely to be active in generating a given

4 document;

5   considering a new term in the given document, wherein the new term is

6 not associated with the cluster; and

7   adding the new link between the cluster and the new term.

1   10.   The method of claim 8, wherein introducing a new link can

2 involve:

3   considering a first cluster that is likely to be active in generating a given

4 document;

5   considering a second cluster that is likely to be active in generating the

6 given document, wherein the second cluster is not associated with the first cluster;

7 and

8   adding the new link between the first cluster and the second cluster.

1   11.   The method of claim 1, wherein producing the new model

2 additionally involves selectively introducing new cluster nodes into the current

3 model.

1   12.   The method of claim 11, wherein selectively introducing a new

2 cluster node involves:

73

3   examining a given document;

4   creating the new cluster node;

5   creating links between the new cluster node and terminals in the given

6 document; and

7   creating links between cluster nodes that are likely to have been involved

8 in generating the given document and the new cluster node.

1 13. The method of claim 1, wherein producing the new model involves

2 calculating an activation for each cluster node in each document, wherein the

3 activation for a given cluster node indicates how many links are likely to fire from

4 the given cluster node to other nodes.

1 14. The method of claim 1, wherein producing the new model involves

2 renumbering clusters in the current model to produce a cluster numbering for the

3 new model; and

4   wherein clusters that are likely to be active in generating more documents

5 are assigned lower numbers that occur earlier in an identifier space, whereas

6 clusters that are likely to be active in generating fewer documents are assigned

7 higher numbers that occur later in the identifier space.

1 15. The method of claim 1, wherein applying a given document to the

2 current model involves:

3   updating a summary variable for each cluster that is likely to be active in

4 the given document, wherein the summary variable summarizes the probabilistic

5 cost of the cluster linking to terminals not existing in the given document; and

6   for terminals that actually do exist in the given document, canceling the

7 effects of corresponding updates to the summary variables.

1   16.   The method of claim 1, wherein applying the set of training
2   documents to the current model involves computing once for each cluster the
3   probabilistic cost of the cluster existing in a document and triggering no words,
4   and for each document applying this cost and subtracting the effects of words that
5   do exist in the document.

1   17.   The method of claim 1, wherein the probabilistic model includes a
2   universal node that is always active and that has weighted links to terminal nodes
3   and/or cluster nodes.

1   18.   A computer-readable storage medium storing instructions that
2   when executed by a computer cause the computer to perform a method for
3   learning a generative model for text, the method comprising:
4   receiving a current model, which contains terminal nodes representing
5   random variables for words and can contain cluster nodes representing clusters of
6   conceptually related words;
7   wherein nodes in the current model are coupled together by weighted
8   links, so that if a cluster node in the probabilistic model fires, a weighted link
9   from the cluster node to another node causes the other node to fire with a
10   probability proportionate to the link weight;
11   receiving a set of training documents, wherein each training document
12   contains a set of words; and
13   applying the set of training documents to the current model to produce a
14   new model.

1   19.   The computer-readable storage medium of claim 18, wherein

2 applying the set of training documents to the current model involves:

3   applying the set of training documents to the links defined in the current

4 model to produce functions for weights for corresponding links in the new model;

5 and

6   optimizing the functions to produce weights for links in the new model.

1   20.   The computer-readable storage medium of claim 19, wherein for a

2 given link, producing functions for a weight on the given link involves:

3   producing a function for the given link for each document in the set of

4 training documents; and

5   multiplying the functions for each document together to produce a

6 function to be optimized for the given link.

1   21.   The computer-readable storage medium of claim 20, wherein for

2 the given link the function for a document is an approximation of the probability

3 of the document's terminals firing as a function of the weight on the given link,

4 keeping all other link weights in the model constant.

1   22.   The computer-readable storage medium of claim 18, wherein the

2 method further comprises iteratively:

3   considering the new model to be the current model; and

4   applying training documents to the current model to produce a subsequent

5 new model.

1   23.   The computer-readable storage medium of claim 22, wherein

2 during an initial iteration, the method further comprises generating an initial

3 current model from a set of words by:

4   generating a universal node that is always active;

5   generating terminal nodes representing words in the set of words; and

6   directly linking the universal node to the terminal nodes.

1   24.   The computer-readable storage medium of claim 22, wherein each

2 iteration uses twice as many training documents as the previous iteration until all

3 available training documents are used.

1   25.   The computer-readable storage medium of claim 18, wherein

2 producing the new model additionally involves selectively introducing new links

3 from clusters to nodes and from clusters to clusters.

1   26.   The computer-readable storage medium of claim 25, wherein

2 introducing a new link can involve:

3   considering a cluster that is likely to be active in generating a given

4 document;

5   considering a new term in the given document, wherein the new term is

6 not associated with the cluster; and

7   adding the new link between the cluster and the new term.

1   27.   The computer-readable storage medium of claim 25, wherein

2 introducing a new link can involve:

3   considering a first cluster that is likely to be active in generating a given

4 document;

77

5   considering a second cluster that is likely to be active in generating the

6   given document, wherein the second cluster is not associated with the first cluster;

7   and

8   adding the new link between the first cluster and the second cluster.

1   28.   The computer-readable storage medium of claim 18, wherein

2   producing the new model additionally involves selectively introducing new cluster

3   nodes into the current model.

1   29.   The computer-readable storage medium of claim 28, wherein

2   selectively introducing a new cluster node involves:

3   examining a given document;

4   creating the new cluster node;

5   creating links between the new cluster node and terminals in the given

6   document; and

7   creating links between cluster nodes that are likely to have been involved

8   in generating the given document and the new cluster node.

1   30.   The computer-readable storage medium of claim 18, wherein

2   producing the new model involves calculating an activation for each cluster node

3   in each document, wherein the activation for a given cluster node indicates how

4   many links are likely to fire from the given cluster node to other nodes.

1   31.   The computer-readable storage medium of claim 18, wherein

2   producing the new model involves renumbering clusters in the current model to

3   produce a cluster numbering for the new model; and

78

4   wherein clusters that are likely to be active in generating more documents

5 are assigned lower numbers that occur earlier in an identifier space, whereas

6 clusters that are likely to be active in generating fewer documents are assigned

7 higher numbers that occur later in the identifier space.

1   32. The computer-readable storage medium of claim 18, wherein

2 applying a given document to the current model involves:

3   updating a summary variable for each cluster that is likely to be active in

4 the given document, wherein the summary variable summarizes the probabilistic

5 cost of the cluster linking to terminals not existing in the given document; and

6   for terminals that actually do exist in the given document, canceling the

7 effects of corresponding updates to the summary variables.

1   33. The computer-readable storage medium of claim 18, wherein

2 applying the set of training documents to the current model involves computing

3 once for each cluster the probabilistic cost of the cluster existing in a document

4 and triggering no words, and for each document applying this cost and subtracting

5 the effects of words that do exist in the document.

1   34. The computer-readable storage medium of claim 18, wherein the

2 probabilistic model includes a universal node that is always active and that has

3 weighted links to terminal nodes and/or cluster nodes.

1   35. An apparatus that learns a generative model for text, comprising:

2   a receiving mechanism configured to receive a current model, which

3 contains terminal nodes representing random variables for words and can contain

4 cluster nodes representing clusters of conceptually related words;

5   wherein nodes in the current model are coupled together by weighted

6   links, so that if a cluster node in the probabilistic model fires, a weighted link

7   from the cluster node to another node causes the other node to fire with a

8   probability proportionate to the link weight;

9   wherein the receiving mechanism is configured to receive a set of training

10   documents, wherein each training document contains a set of words; and

11   a training mechanism configured to apply the set of training documents to

12   the current model to produce a new model.